Dean&Francis
ISSN 2959-6130

Regression and Classification Approaches to Microsoft Stock Forecasting with Machine Learning

Shuofeng Song

[1]Department of Arts and Sciences, The Ohio State University, Zhengzhou, China
*Corresponding author: songshuofeng@gmail.com

Abstract:

This study discusses the application of machine learning algorithms for the prediction of Microsoft stock prices using historical data for five years. Preprocess of the data was done by treating missing values, creating lag features, and normalizing the data for better model performance. To the price of close and momentum, various models were trained, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression, and Gradient Boosting Regressor. Based on the experimental results, Linear Regression achieved the best performance in closing price prediction, recording a Coefficient of Determination (R^2) of 0.91, Mean Squared Error (MSE) of 45.77, Mean Absolute Error (MAE) of 4.64, and Mean Absolute Percentage Error (MAPE) of 0.01. For momentum prediction, Linear Regression again outperformed other models, achieving $R^2 = 0.51$, MSE = 21.25, MAE = 2.79, and MAPE = 2.05. And other models showed much weaker explanatory power. When predicting the Relative Strength Index (RSI) classification, Gradient Boosting delivered the best overall performance, achieving Accuracy = 1.00, F1 score = 1.00, Cross-Validation (CV) mean accuracy = 0.998, and CV standard deviation (CV std) = 0.02. Although other models such as Linear Regression, Logistic Regression, Support Vector Classifier, and Random Forest achieved strong results, none matched the superior performance of Gradient Boosting.

Keywords: Stock Price Prediction; Machine Learning; Regression Models; Ensemble Learning; Microsoft Stock.

1. Introduction

Over the past decade, data-driven technologies have achieved exponential growth, which has transformed the way investment analysis is conducted and the model of financial decision-making. Among all the methods for building models for financial issues and extracting useful insights from vast amounts of data,

machine learning is very crucial.

In recent years, the application of machine learning to stock price prediction has made significant progress. Albert Wong published two notable studies: the first employed technical data and exogenous variables (such as stock market indices, interest rates, and inflation indicators) to conduct high-frequency prediction experiments, updating data every 15 minutes and developing several simple algorithms. The results showed that shorter prediction horizons significantly reduced errors—at a 15-minute interval, Extreme Gradient Boosting (XGBoost) and Random Forest achieved Root Mean Square Error (RMSE) values of about 12–14 USD and Mean Absolute Percentage Error (MAPE) below 1% [1]. Each training and prediction cycle took less than 2 seconds, allowing a full week of 15-minute interval forecasts to be generated in under 3 minutes. For one-day forecasts, XGBoost still performed best, with Tesla's RMSE as low as about 8 USD and MAPE controlled within 5% [1]. Another study compared four machine learning algorithms for short-term predictions of Tesla, Apple, and Nvidia stock prices, confirming that incorporating macroeconomic variables such as interest rates, gold, and oil significantly improved model stability and predictive accuracy [2]. Meanwhile, Yao combined neural networks and Support Vector Machine (SVM) for feature engineering and prediction, finding that Long Short-Term Memory (LSTM) excelled in accuracy with high-dimensional, nonlinear data, while SVM offered faster training; they recommended constructing a hybrid LSTM–SVM model [3]. Archit further validated that neural networks, deep learning, and hybrid models (e.g., Autoregressive Integrated Moving Average (ARIMA)–SVM, Support Vector Regression (SVR)–Artificial Neural Network (ANN)–RF enhance predictive power when integrated with sentiment and macroeconomic factors, achieving a directional accuracy of 70.59% when using only news sentiment to forecast stock price trends [4].

In terms of traditional methods and ensemble strategies, Shailendra Gaur found that SVM achieved the highest accuracy in trend prediction, with Boosting providing further improvements [5]. Huang, using 22 years of quarterly financial data from S&P 100 constituents, confirmed that Random Forest was the top performer, with an average quarterly excess return of 1.63% and a cumulative 18-quarter return of 33.5%; multi-model Bootstrap Aggregation achieved an even higher cumulative return of 137% [6]. Sentiment analysis also proved important. Kompella combined news polarity scores with Random Forest, outperforming logistic regression across Mean Absolute Error (MAE), Mean Squared Error (MSE), and explained variance metrics [7]. Obthong's survey reported that Naïve Bayes and RF could reach accuracies of up to 90%, sentiment-enhanced SVM up to 60%, and deep learning methods—especially LSTM/ Bidirectional Long Short-Term Memory (BLSTM)—achieved short-term prediction Root Mean Square Error (RMSE) as low as 0.00947 [8]. On the other hand, Zhang showed that historical curve-shape features had no significant impact on future returns, highlighting the need to develop new methods to capture Non Curve Shape Features (NCSF) [9]. Addressing single-model bias, Wei proposed an improved ensemble model combining SVR, linear regression, and K-Nearest Neighbors (KNN), reducing Amazon's RMSE from 0.0685 to 0.0318 and increasing R^2 to 0.90, while lowering Tian Chang Group's RMSE from 273.84 to 82.43 and raising R^2 to 0.96, with 10-fold cross-validation scores close to 0.9, underscoring the robustness and high accuracy of ensemble learning [10].

This study uses five years of Microsoft stock data (June 2020–June 2025) to predict closing price, momentum, and Relative Strength Index (RSI). Engineered features feed linear, nonlinear, and ensemble models, whose performance is assessed by Coefficient of Determination (R^2), MSE, MAE, MAPE for price and momentum, and by accuracy, F1, and cross-validation metrics for Relative Strength Index (RSI) to identify the most effective approach for investment analysis and stock price forecasting.

2. Methodology

2.1 Dataset

The data employed in this study contains the stock market trading history of Microsoft Corporation from 2020 to 2025, with 1,256 records. (Saman Fatima, Microsoft Stock Price Data (Last 5 Years) [11]. It possesses six features: Opening Price (Open), Closing Price (Close), Highest Price (High), Lowest Price (Low), Trading Volume (Volume), and Transaction Date (Date). They are all used to provide a general description of the change in the stock market of Microsoft over time and are the foundation for applying machine learning models to predict stock prices.

2.2 Algorithm

This paper adopts a method framework integrating data mining, feature engineering, regression analysis, and time series analysis. Discover the hidden patterns in historical stock market data and draw some insights that can provide useful information for investment decisions.

To achieve this goal, this paper employs several regression algorithms on the processed dataset, including Linear Regression, which models the relationship between features and target as a straight line and offers simplicity and interpretability; Logistic Regression, a probabilistic linear model that can handle classification-like boundaries and is robust for binary outcomes; Decision Tree Regression,

ISSN 2959-6130

which partitions data into hierarchical rules and captures nonlinear relationships without requiring feature scaling; Random Forest Regression, an ensemble of multiple decision trees that reduces variance and improves generalization through bagging; Support Vector Regression and Support Vector Classification, which uses kernel functions to map data into higher-dimensional spaces for capturing complex nonlinear patterns while maintaining strong regularization; and Gradient Boosting Regression, which builds an additive model of sequential trees to minimize errors, excelling in accuracy and handling complex feature interactions.

After that, the model performance of closing price and momentum price would be evaluated by R^2, MSE, MAE, and MAPE. For the prediction of the Relative Strength Index (RSI), this paper adopts a classification-based evaluation criterion, which includes Accuracy, F1(macro), CV mean accuracy, and CV std.

3. Results

3.1 Model Performance

3.1.1 Price of Close

Table 1. Performance Comparison of Machine Learning Models for Microsoft Closing Price Prediction

	R^2	MSE	MAE	MAPE
Linear Regression	0.91	45.77	4.64	0.01
Decision Tree	0.84	80.64	7.03	0.02
Random Forest	0.89	54.34	5.44	0.01
SVR	-28.33	14758.85	119.24	0.29
Gradient Boosting	0.87	57.16	5.46	0.01

In this study, several machine learning models are employed to predict Microsoft's stock prices, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), and Gradient Boosting. By comparing the R^2 score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), the strengths and weaknesses of each model can be clearly observed.

The results from Table 1 point to Linear Regression having the best performance with an R^2 of 0.91, lowest MSE (45.77), MAE (4.64), and MAPE (0.01). This implies that it provides the greatest prediction accuracy and reliability. Random Forest and Gradient Boosting also performed well with R^2 of 0.89 and 0.87 respectively. Their error metrics were within tolerable limits, exhibiting great fitting ability and generalization performance. On the other hand, the Decision Tree model was significantly weaker, with significantly higher errors compared to the best models.

SVR showed the poorest performance, with an R^2 of -28.33, far below the other models. Its MSE and MAE were extremely high (14,758.85 and 119.24), and the MAPE reached 0.29. These results indicate that SVR had very weak predictive power on this dataset and cannot be considered an effective model.

Taking all metrics into account, Linear Regression is the most effective model in this study, followed by Random Forest and Gradient Boosting. These models demonstrated superior predictive accuracy and generalization ability, making them suitable for Microsoft stock price prediction.

3.1.2 Momentum

Table 2. Performance Comparison of Regression Models for Microsoft Stock Momentum Prediction

	R^2	MSE	MAE	MAPE
Linear Regression	0.51	21.25	2.79	2.05
Decision Tree	0.12	37.73	4.29	290.92
Random Forest	0.25	32.21	4.05	243.46
SVR	-8.91	42.96	4.48	1.03
Gradient Boosting	0.17	35.82	4.35	2.70

In this part, the same models and evaluation criteria used for closing price prediction were applied to assess performance. The results from Table 2 indicate that Linear Regression worked best among the models with the highest R^2 at 0.51, and also the lowest MSE (21.25), MAE (2.79), and MAPE (2.05). It demonstrates its relatively high pre-

dictive accuracy and validity in identifying patterns of stock momentum. Random Forest and Gradient Boosting exhibited middle-of-the-road performances with R^2 of 0.25 and 0.17, but their higher MSE and MAPE values reflect worse generalization compared to Linear Regression.

The Decision Tree model was also low in terms of prediction, with a low R^2 of 0.12 and a very high MAPE (290.92), demonstrating predictive instability. SVR was the worst performer in that it demonstrated a negative R^2 value of -8.91 and the highest error rates in both MSE (42.96) and MAE (4.48), although possessing a low MAPE (1.03). This is evidence that SVR did not capture the momentum dynamics appropriately. Overall, Linear Regression is the best model to use for momentum prediction, with Random Forest and Gradient Boosting coming second. Decision Tree and SVR were not good at predicting and should not be used for good momentum prediction in this case.

3.1.3 RSI

Table 3. Performance Comparison of Machine Learning Models for Microsoft RSI Prediction

	Accuracy	F1	CV mean accuracy	CV std
Linear Regression	0.86	0.31	0.24	0.10
Logistic Regression	0.93	0.60	0.97	0.01
Random Forest	0.97	0.66	0.97	0.05
SVC	0.86	0.31	0.92	0.0018
Gradient Boosting	1	1	0.998	0.02

More machine learning models were also employed to predict Microsoft's RSI values, and those are Linear Regression, Logistic Regression, Random Forest, Support Vector Classification (SVC), and Gradient Boosting. Through the comparison of Accuracy, F1 score (macro-averaged), CV mean accuracy, and CV std, the comparative merits and demerits of each model can be clearly observed.

The result of Table 3 shows that Gradient Boosting worked best with the highest CV mean accuracy (0.998) and lowest CV std (0.02), which suggests high predictive ability and stability. Random Forest was also good with a CV mean accuracy of 0.97 with relatively low CV std (0.05), which proved its potency. Logistic Regression provided decent results with a CV mean accuracy of 0.97, but slightly lower F1 than Random Forest.

Conversely, Linear Regression performed less optimally in RSI prediction. Although it also possessed a relatively high accuracy score of 0.86, its weak generalizability, as seen from its low CV mean accuracy value of 0.24, renders it less reliable compared to other models.

According to all the metrics, Gradient Boosting is the best-performing among the models in RSI prediction, followed by Random Forest and Logistic Regression. These three were highly predictive and strong, therefore being the best possible approach to RSI forecasting under this work.

4. Discussion

The performance differences observed across tasks arise primarily from the statistical characteristics of the target variables. The closing price represents a relatively stable time series with a clear trend, low noise, and strong linearity, making it well-suited to traditional linear regression models. Momentum, which captures the rate of price change, is far more volatile, highly nonlinear, and noisier, reducing R^2 values and complicating the detection of stable patterns. In contrast, the Relative Strength Index (RSI) is a bounded technical indicator used for classification; its discrete output (e.g., overbought/oversold) differs greatly from continuous regression targets and favors models capable of handling nonlinearity and boundary separation, such as Gradient Boosting and Random Forest. Data distribution and signal-to-noise ratio further accentuate these differences: closing prices typically exhibit strong trends and autocorrelation with high signal and low noise; momentum is heavily affected by short-term random fluctuations, resulting in a low signal-to-noise ratio; and RSI, derived through multi-step calculations with smoothing, has reduced noise, making it especially compatible with classification models.

5. Conclusion

This study utilized five years of Microsoft's stock data and applied machine learning techniques to predict closing prices, momentum, and Relative Strength Index (RSI) values. After evaluating multiple models in both regression and classification tasks, distinct performance patterns emerged. For closing price prediction, Linear Regression achieved the highest accuracy and stability, followed by Random Forest and Gradient Boosting, both of which showed strong generalization capabilities. In momentum prediction, Linear Regression again outperformed other models, while Decision Trees and Support Vector Regres-

ISSN 2959-6130

sion (SVR) lagged behind. For RSI prediction, Gradient Boosting delivered the best cross-validation accuracy and stability, with Random Forest ranking second.

The performance differences across these tasks stem largely from the statistical characteristics of the target variables. The closing price is a relatively stable time series with a clear trend, low noise, and strong linearity, making it well suited to linear models. Momentum, which measures the rate of price change, is highly volatile, strongly nonlinear, and noisy, lowering R^2 and complicating pattern detection. RSI, as a bounded technical indicator used for classification, outputs discrete categories (e.g., overbought/oversold) and thus favors models adept at handling nonlinearity and boundary separation, such as Gradient Boosting and Random Forest. Differences in data distribution and signal-to-noise ratio reinforce these outcomes: closing prices exhibit strong trends and high signal with low noise; momentum suffers from short-term fluctuations and a low signal-to-noise ratio; and RSI, derived through multi-step calculations with smoothing, has reduced noise and is especially compatible with classification models.

Despite these insights, this research has limitations. The five-year dataset may not fully capture structural shocks in the market, and the scope of feature engineering is relatively narrow. Future research can expand this framework by developing a more comprehensive hybrid architecture, leveraging automated machine learning for model optimization, and performing out-of-sample testing to ensure scalability and stability in real-world deployment.

References

[1] Albert W, Juan F, Raheem A, et al. Forecasting of stock prices using machine learning models. IEEE Systems Conference, 2023, 20(15): 1–7.

[2] Albert W, Steven W, Emilio S, et al. Short-term stock price forecasting using exogenous variables and machine learning algorithms. 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), 2023, 56(19): 260–265.

[3] Wen Q Y, You W, Zhang S F, Chang E C, et al. Stock price analysis and forecasting based on machine learning. Conference on Computer Science and Communication Technology, 2022, 29(12): 1250660–1250668.

[4] Archit A V, Paresh J T. A survey of machine learning techniques used on Indian stock market. IOP Conference Series: Materials Science and Engineering, 2021, 36(10): 236–239.

[5] Gaur S, Bhardwaj R, Bansal V, et al. Stock market prediction using machine learning. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 2019.

[6] Huang Y, Capretz L F, Ho D. Machine learning for stock prediction based on fundamental analysis. IEEE Symposium Series on Computational Intelligence, 2021: 1–10.

[7] Kompella S, Chilukuri K C. Stock market prediction using machine learning methods. International Journal of Computer Engineering & Technology, 2019.

[8] Obthong M, Tantisantiwong N, Jeamwatthanachai W, et al. A survey on machine learning for stock price prediction: algorithms and techniques. International Conference on Finance, Economics, Management and IT Business, 2020: 63–71.

[9] Zhang P, Yang J Y, Zhu H, et al. Failure in stock price prediction: a comparison between the curve-shape-feature and non-curve-shape-feature modes of existing machine learning algorithms. International Journal of Computers Communications & Control, 2021, 16.

[10] Wei Z, Chen Y, Gao M, et al. Stock prediction methods based on ensemble learning. Academic Journal of Business & Management, 2021.

[11] Fatima S. Microsoft stock price data (last 5 years). Published 2023-07-19. Accessed 2025-09-16.